Virtual Microdata Laboratory User Guide

1. Introduction

This document aims to set out an overview of the Virtual Microdata Laboratory (VML) and highlight some key features of best practice for users.

The VML is a secure environment for research access to microdata, which is managed by the VML team within the Office for National Statistics (ONS). The use of anonymised, but potentially disclosive, survey returns requires a strict adherence to legal and technical procedures. Failure to follow proper working practice may result in the removal of access for researchers and sanctions for their parent organisation.

This document is intended as a guide for users. It aims to highlight the following:

- The locations where the VML can be accessed, and how to book a time slot.
- Security procedures necessary to obtain access, and security procedures which are in place while work is conducted in the VML.
- The structure of the VML - how to place files in the laboratory and what software is available for research/analysis.
- How to use the VML - logging in and best practice within the laboratory.
- Disclosure control - how research output should be constructed, timescales for output clearance, and Intermediate vs. Final Disclosure clearance.

2. Accessing the Laboratory

2.1 Locations

The VML can be accessed from the following locations:

London	Newport	Southport	Titchfield
1 Drummond Gate Pimlico London SW1V 2QQ (8 terminals)	Cardiff Road Newport South Wales NP10 8XG (3 terminals)	Smedley Hydro Trafalgar Road Birkdale Southport PR8 2HH (1 terminal)	Segensworth Road Titchfield Fareham PO15 5RR (1 terminal)

2.2 Booking time in the VML

Users must first complete the VML researcher training course, and a research application form to gain approval to use the VML. The application form should cover an overview of the aims of the research project and which datasets you plan to use.

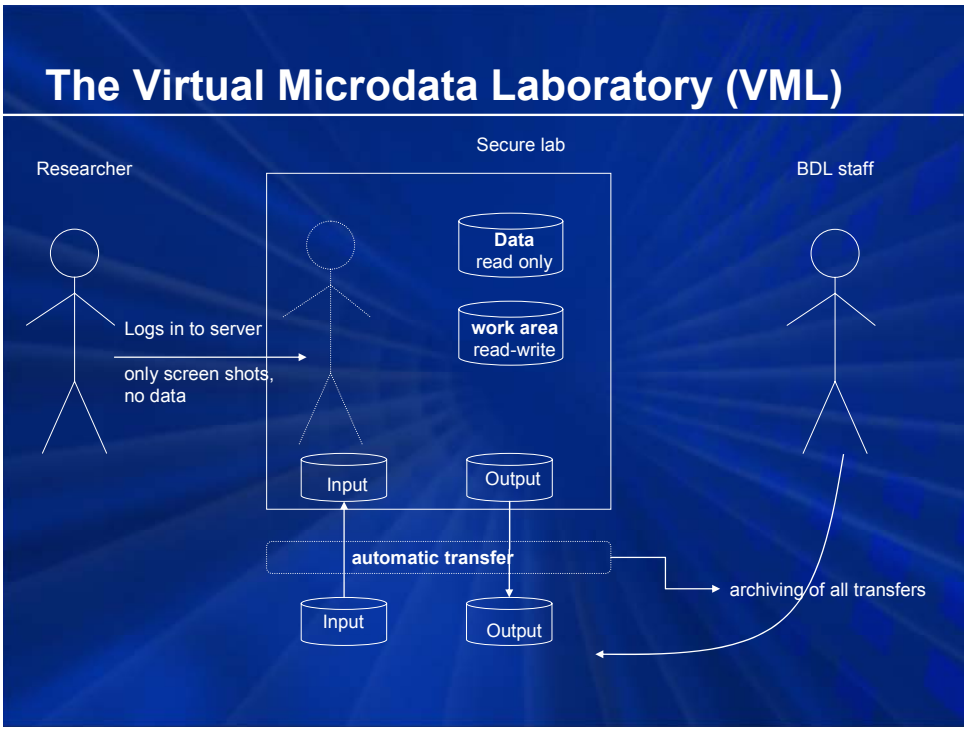
Once your application has been cleared a booking must be made to use the VML. The VML is a shared resource so its use must be planned accordingly. It is not possible for users to just arrive at one of the laboratory locations – access will be denied. The VML team must always consider the security issues of letting users into ONS sites and the confidential nature of the data. To book a place on the training course, request an application form or book a time slot in the VML email: vml@ons.gov.uk.

The VML is open 9:30am to 4:30pm on working days.

2.3 What to expect

Within the VML access to the data sets is through "thin terminals". These are essentially a network card attached to a screen and a keyboard. They will only connect to the VML server; users are cut-off from all outside sites and servers.

The VML is structured as follows:



The diagram above illustrates the basic structure of the VML. It should be noted that only VML staff can transfer data in and out of the VML. Users have read-only access to the data sets, and read-write abilities in the work areas. VML staff provide a limited query-driven response service to users.

3. Security

3.1 Clearance procedures to access the VML

To protect the confidentiality of the data, access is strictly controlled and only researchers covered by a secondment contract can access the Lab.

All potential users must complete the VML training course (1-2 hours). This is provided to give knowledge of expected working procedures. If adhered to it will speed up the release of research outputs.

Users will be allocated a risk assessment score by the VML. This will determine whether users will be granted access to the VML and how strict the output clearance procedures will be. Users risk score primarily depends on the following factors:

- Credibility of research background.
- Evidence of trustworthiness (e.g. previous experience of VML, number of formal warnings, previous breaches etc)
- Possibility for conflict of interest.

3.2 Security Procedures within the VML

Inside the Lab it is not possible to access the printer, floppy disk, CD-ROM, Internet or the rest of the ONS network. Paper is allowed to be used to make notes when in the VML, but researchers must take care to guard the confidential nature of anything written down. Written notes must be destroyed at the end of each working day. The use of laptops or personal organisers (e.g. Blackberry) is not allowed.

Although once inside the VML it is possible to access all the available datasets, researchers should only access those specified in the secondment contract for which permission has been granted. When working with the datasets researchers must not attempt to identify a single company or individual.

Given the confidential nature of the data, under no circumstances should raw data be removed from the VML. Analytical output or statistical tables may only be removed from the VML and sent to researchers once they have gone through disclosure control by VML staff members. All transfers of data to and from the VML will be recorded.

As researchers cannot accidentally remove data from the lab, all unauthorised releases are deliberate breaches of contract and action will be taken accordingly. Attempting to remove uncleared data (whether that attempt is successful or not) is grounds for immediate termination of the secondment and the removal of the researcher from the premises. There is also the potential for the removal of access rights for all employees of the supporting institution and criminal charges against the researcher.

3.3 The "safe" framework

The set-up of the Lab has been designed to make it very difficult for users to break the legal and practical controls that have been put in place. However, it is not possible to make the VML 100% safe. The VML team places an emphasis on trusting its approved researchers; the Lab has therefore been designed to operate within a safe projects/safe settings/safe people/safe outputs framework. Breaches of trust will set in motion security review procedures and disciplinary actions.

4. Structure of the Laboratory

4.1 The Directories

If users require files in addition to the data sets available in the VML (e.g. for data linking/fusion), these should be emailed to the VML team who will then place the files in one of the directories. If additional files contain confidential data users should speak to the VML team about the possibility of setting up a separate version of the VML for restricted access. Users may also consider setting up passwords on their Excel spreadsheets.

Output files can only be removed from the VML by VML staff. When research has been completed it must go through the disclosure control procedures outlined in section 6 (and the training course).

4.2 Overview of the datasets

An overview of the data sets contained in the VML is available in a separate word document entitled "VML Data Sets". It aims to highlight the latest data available and any associated documentation that is available for potential users. This can be requested from the VML team, and may be useful for the writing of application forms.

4.3 Software available within the VML

The following software packages are available for use within the VML:

- Microsoft Excel
- Microsoft Word
- Microsoft PowerPoint
- Stata 9
- SPSS 12
- SAS
- Gauss 5
- SuperCROSS
- StatTransfer 7
- WinZip
- Acrobat Reader
- ArcCatalogue
- ArcMap

- ArcReader

There are some specialist Stata programme files (.ado) stored in the VML. They can be found in the documentation drive (X:\code\ado). If you require a particular (new) specialist Stata programme please download it and email the .ado file as an attachment to vml@ons.gov.uk we will then place it in the VML documentation drive for your use.

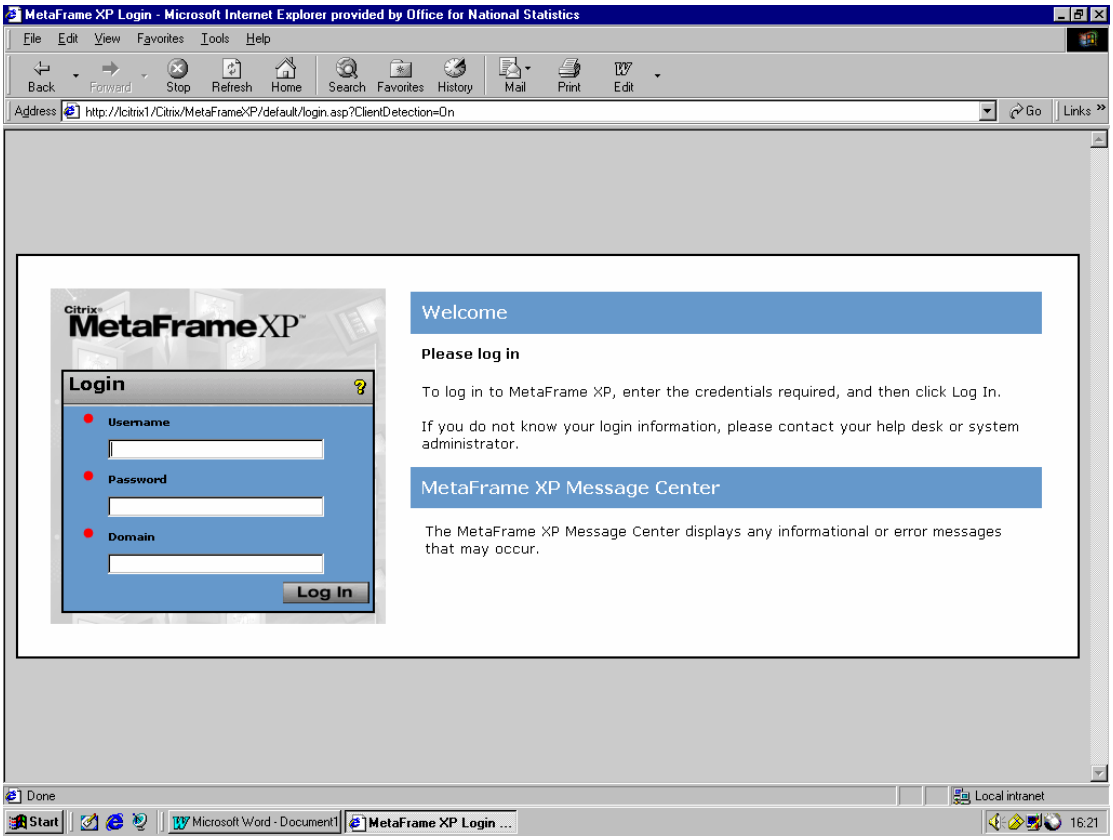
Specialist software requirements are dealt with by request. It may not be possible to provide other software due to licensing agreements and costs of provision.

5. VML use

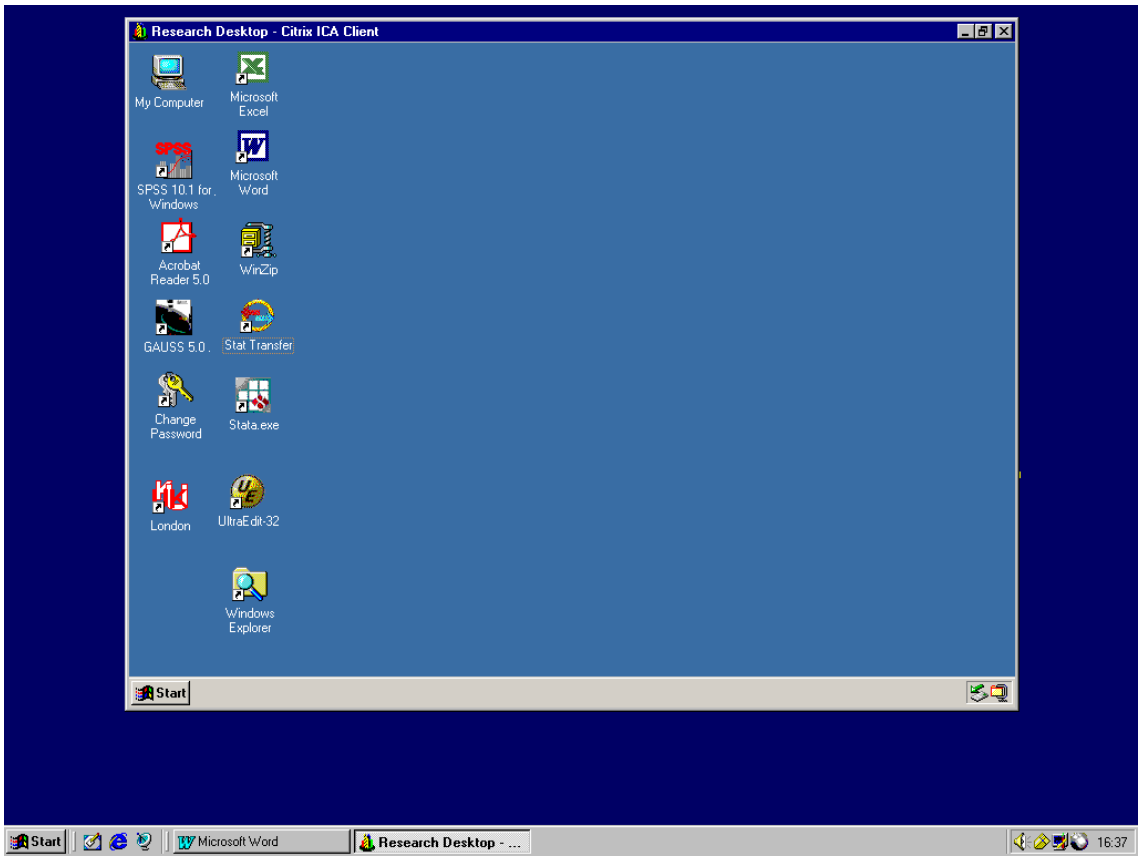
5.1 Logging in to the VML

Researchers will be issued a username and password from the VML team to allow access to the VML. The following outlines the steps that should be taken to log in:

- From a PC: launch internet explorer and type http://lcitrix1 this will bring up the CITRIX web interface.
- From a Thin Ternimal: switch on the terminal and the log in screen will appear automatically. The screen will appear as below. Always ensure that the Domain is set to London (no matter which site you are accessing the VML from).


MetaFrame XP Login - Microsoft Internet Explorer provided by Office for National Statistics
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://lcitrix1/Citrix/MetaFrameXP/default/login.asp?ClientDetection=On
Go Links
Citrix MetaFrameXP
Login
Username
Password
Domain
Log In
Welcome
Please log in
To log in to MetaFrame XP, enter the credentials required, and then click Log In.
If you do not know your login information, please contact your help desk or system administrator.
MetaFrame XP Message Center
The MetaFrame XP Message Center displays any informational or error messages that may occur.
Done
Local intranet
Start
Microsoft Word - Document1
MetaFrame XP Login ...
16:21

- Once you have entered your username and password you will be directed to the research desktop window. See below:



- If you are experiencing problems while using the VML (or forget your password etc) please contact the VML administrator on (72) 5844.
- If you require additional information about the data sets please email vml@ons.gov.uk.

5.2 File management and best practice

While using the VML please bear in mind that it is a shared resource. If you claim excessive amounts of memory to run your programmes and/or keep multiple copies of the same data sets, the server will run more slowly and it will take longer to complete your work. This problem is compounded the more people do it, and is analogous to the "tragedy of the commons" theory.

We recommend that researchers abide by the following guidelines:

- Only claim memory sufficient to run programmes.
- Be careful when merging/appending files – they may become very large.
- Drop irrelevant variables where ever possible - use the "keep" and "collapse" options to leave only essential data.
- Delete unwanted files and any copies at the end of your daily sessions.

- Compress data files if you are not returning to the VML in the near future (e.g. use WinZip).
- Ask the BDL team to run large production jobs overnight. This will be a much more efficient use of your time, and will stop the VML slowing down for other users.
- Save your work regularly while working in the VML. On rare occasions the VML has experienced severe technical difficulties.

6. Disclosure Control

Before researchers' work can leave the VML it must go through disclosure control by the VML team to ensure that output does not contain any confidential data, or that confidential information could be calculated through inference.

6.1 Rules for disclosure control

The following is a summary of the disclosure control rules as taught in the training course:

- Threshold rule: cells backed with less than 10 units will not be cleared. (Local unit analysis e.g. BSD must show that cells are also backed up by at least 3 enterprise groups.)
- Dominance rule: the sum of all but the largest two units must be greater than 10% of the largest unit.
- Primary disclosure: no maximum or minimum values are to be included unless they can be shown to be uninformative. Graphical output must include the frequencies which underpin each data point.
- Secondary disclosure: can occur from differencing across tables, and subtracting from totals and zero values in cells. Researchers should consider whether it is appropriate to transform tabulated data to ensure that it is non-disclosive.

There are not rules to cover every eventuality in the Lab. If you have questions regarding your output please contact a member of the VML team.

6.2 Structure of Outputs

The VML team will only clear files for disclosure if users' files are made up of clean output. For example, we do not want to see files which contain the construction of variables, tables and regressions that went wrong and then had to be re-done later, errors within Stata logs etc. If researchers do not present their output in a clean format, it will be returned to them and changes must be made before it will be cleared.

An excessive quantity of outputs will result in a rejection of output clearance. It is the researchers' responsibility to provide the VML team with an appropriate length of output.

Tables and regressions should be placed in separate output files for clearance. Users must always provide the frequencies underlying any graphs which are put

forward for clearance. Generally, regressions are less time consuming for VML staff to clear. Researchers are advised to note that hiding coefficients makes linear and non-linear regressions completely non-disclosive, while panel models are inherently safe. Presenting work in these formats will considerably speed-up the clearance process.

6.3 Intermediate vs. Final Disclosure

To receive data files for publication and public use output must be cleared through the Final Disclosure procedure. Output must be cleared by two members of VML staff; hence, researchers should factor in more time to get work cleared to this level. Overall, output must be of a high standard and meet the disclosure rules in its strictest forms.

Intermediate clearance is not for public use. Files given this level of clearance may only be used for writing up results, recoding programmes and discussion with co-researchers. Intermediate output may not be used for publication; which includes presentations at seminars and workshops, even if no paper is circulated with the presentation. The Intermediate Disclosure procedure needs only to be cleared by one member of VML staff; hence, is generally faster.

6.4 Status of outputs

All the outputs are ONS products – so if researcher A runs a regression of X on Y in 1997 on a specific subset of the data, and Researcher B asks ONS for the same regression, they must be given the results under Freedom of Information and the National Statistics Code of Practice. This is why the results are ONS copyright.

It is of course unlikely that someone not involved in a project would ask for exactly the same results, but not impossible. For example, a paper published in a journal may have a statement like "other analyses showed similar results", and ONS may be asked for these by an interested reader. In the first instance ONS would normally ask the interested researcher to get in touch with the author, as one aim of the VML is to develop an effective research community.

6.5 Timescale for clearance

Researchers should expect VML staff to take up to 5 working days to clear their output and should plan accordingly. Although the VML team will try to clear output quickly, it is not realistic to expect a quick turn around time from the VML staff on all occasions. The VML team will not re-arrange their work schedule to speed up clearance so that researchers can present output at conferences etc. It is in users' interests to take into account UK public holidays when planning their work schedule.

Users should always email vml@ons.gov.uk for clearance requests not just individual team members. It may be appropriate to "cc" individual team members if they have cleared previous versions of the output, but ensure that the VML email address is the main point of contact. This will ensure that your response is dealt with when individual team members are on leave.